IDCC14 | *Practice Paper*

Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping

Herbert Van de Sompel
Los Alamos National Laboratory

Robert Sanderson
Los Alamos National Laboratory

Harihar Shankar
Los Alamos National Laboratory

Martin Klein
Los Alamos National Laboratory

Abstract

Persistent IDentifiers (PIDs) such as DOIs, Handles, ARK identifiers play a significant role in the identification of a wide variety of assets that are created and used in scholarly endeavours, including research papers, datasets, images, etc. Motivated by concerns about long-term persistence, among others, PIDs are minted outside the information access protocol of the day, HTTP. Yet, value-added services targeted at both humans and machines routinely assume or even require resources identified by means of HTTP URIs in order to make use of off-the-shelf components like web browsers and servers. Hence, an unambiguous bridge is required between the PID-oriented paradigm that is widespread in research communication and the HTTP-oriented web, semantic web and linked data environment. This paper describes the problem, and a possible solution towards defining and deploying such an interoperable bridge.

Introduction

Although the Web and HTTP were initially conceived as conduits for research communication, concerns about the long-term brittleness of HTTP URIs were voiced as soon as scholarly literature started to find its way online. An HTTP URI simultaneously identifies a resource on the web and provides access to it. When a resource moves from one Web location to another, for example because a journal changes ownership, access based on the original HTTP URI ceases. HTTP redirects (HTTP "301 Moved Permanently") are commonly used on the web to address this issue. But, in an environment where publishers are acquired by others and cease to exist, this technique was not adopted. Similar concerns existed outside of the realm of research communication. As a result, approaches were introduced in which identifying is decoupled from locating and access. This paper does not intend to provide a

Submitted 23 October 2013

Correspondence should be addressed to Herbert Van de Sompel, Email herbertv@lanl.gov

comprehensive overview of solutions with that regard; it suffices to illustrate two broad types of approaches distinguished by the nature of the identifier assigned to an asset. The characterization is summarized in Table 1:

- *PURLs*: Under the PURL paradigm, less common in research communication but popular for linked data ontologies, an asset is identified by a special HTTP URI - HTTP-URI-PURL - that is resolved by a PURL server. The PURL server provides a mapping between each HTTP-URI-PURL and an actual network location, HTTP-URI-LOC, and requests for a specific HTTP-URI-PURL are redirected to that network location. When a resource moves on the network, its HTTP-URI-PURL remains stable. The intervention of that HTTP-URI-PURL's administrator ensures that the new network location ends up in the mapping, and subsequent requests are redirected to this new location. In many uses of the PURL paradigm, the resource at HTTP-URI-LOC provides the content for the resource identified by the HTTP-URI-PURL. For example, dereferencing the HTTP-URI-PURL <http://purl.org/dc/terms/> yields the HTTP-URI-LOC <http://dublincore.org/documents/2012/06/14/dcmi-terms/>, a page that defines the Dublin Core Metadata Terms ontology.

- *PIDs*: Under the PID paradigm, made widely popular in scholarship by DOIs, an asset is assigned a non-HTTP identifier, PID. That identifier is typically a string that complies with a well-defined syntax, minted in a namespace controlled by a naming authority. In order to interact with an asset identified by a PID, the PID is used in the path component of an HTTP URI, HTTP-URI-PID. That HTTP URI has a hostname controlled by an authority that is able to do something meaningful with it. The main functionality provided under this paradigm is redirection from that PID-carrying HTTP URI to a network location associated with the asset identified by the PID. Similar to the PURL paradigm, it is powered by a mapping between the identifier of the resource (PID) and the HTTP URI of the network location (HTTP-URI-LAND), and is kept up to date through an administrative process. However, services other than basic redirection have been defined for HTTP-URI-PIDs, including access to metadata about the resource identified by the PID. In most uses of the PID paradigm, the resource at HTTP-URI-LAND does not provide the content for the resource identified by the PID but rather is a descriptive landing page pertaining to the resource; the content itself resides at another HTTP URI, HTTP-URI-LOC, somehow linked from HTTP-URI-LAND. For example, the asset with PID <doi:10.1145/1998076.1998111> with HTTP-URI-PID <http://dx.doi.org/10.1145/1998076.1998111> has HTTP-URI-LAND <http://dl.acm.org/citation.cfm?id=1998076.1998111>, which is the URI of a landing page for human consumption. The PDF content for this asset is at HTTP-URI-LOC <http://dl.acm.org/ft_gateway.cfm?id=1998111&type=pdf>.

	PURL paradigm	PID paradigm
Resource identifier	HTTP-URI-PURL	PID
Resolving URI	HTTP-URI-PURL	HTTP-URI-PID
Redirect URI	HTTP-URI-LOC	HTTP-URI-LAND
Location URI	HTTP-URI-LOC	HTTP-URI-LOC

Table 1: Identification and location under two common paradigms

This paper focuses on the PID paradigm because of its omnipresence in research communication. Initially used to identify journal articles, PIDs are now also deployed or considered for the identification of a wide variety of assets used or created in scholarship including datasets, images, tables, software, peer-reviews and so on. While some guarantee of long-term persistence of linkages within the scholarly record remains a major motivator for assigning PIDs, the emergence of article-level metrics and alt-metrics has brought about a new motivator grounded in the desire to assign academic credit for a variety of web-based scholarly contributions, not just for journal articles.

PIDs and HTTP URIs: The Need to Map

The use of PIDs for research communication is both understandable and justifiable. Yet because the scholarly record, and increasingly the scholarly process, is a long way towards becoming fully web-based, all value-added services provided for PID-identified assets require resources identified by means of HTTP URIs. Hence, an unambiguous bridge is required between the PID-oriented paradigm and the HTTP-oriented web, semantic web and linked data environment.

Interestingly, more than a decade into the use of PIDs in scholarship, such a bridge has not been specified beyond the mere understanding that appending a `PID` to an appropriate base URL and dereferencing the resulting `HTTP-URI-PID` yields a redirection to information pertaining to the PID-identified asset. As described, in most cases the redirection is to a landing page, in other cases it is to an HTML rendering of the asset itself. The obtained information is generally targeted at human consumption. In the case of DOIs, access to machine-actionable, descriptive RDF metadata has recently been added by introducing content negotiation with `HTTP-URI-PID`. This is a most welcome nod to machine agents but there is a need to move beyond the status quo and devise a bridge that takes into account the following significant considerations:

- Increasingly, machine agents consume web resources. A recent study found that 61% of all web traffic is generated by machine consumption. This trend, observed for the web at large, is also visible in web-based scholarship as exemplified by organized machine-oriented access to journal articles for text mining purposes (e.g. CrossRef Prospect).
- Increasingly, assets created in scholarship consist of multiple resources, not just a single component. Where a PID initially identified a PDF article, for example, it increasingly is used to identify a bundle of resources associated with a research endeavour, ranging from multiple renderings of the same article to a dataset with several constituent components, or the union of highly heterogeneous assets used or generated in the course of an experiment. In terms of Table 1, this means that multiple `HTTP-URI-LOCs` correspond with a single `PID`.
- Many resources used or created during the research process (e.g. software, workflows, ontologies) do not have the sense of fixity that resources generated as the end result of that process (e.g. journal articles, books) do. The dynamic nature of the resources used during the research process requires a sensible versioning approach and associated functionality to support discovery and access to resource

versions. In addition, the dynamic and interdependent nature of these resources creates the need for the ability to determine what the state of these resources was at particular moments in their lifecycle. This need was argued at length in The Web as Infrastructure for Scholarly Research and Communication, a keynote at IDCC 2013, and From the Version of Record to a Version of the Record, a keynote at the CNI Spring 2013 meeting. The need is related to the desire for reproducible research explored - among others - in the Wf4Ever project, and its importance is exemplified by the fact that versioning, along with identity, aggregation, provenance, and annotation, is a core ingredient of Research Objects, containers of experimental resources that are essential to a computational scientific study or investigation.

The following, generic, desirable functionalities aptly illustrate the need to devise a *machine-actionable* bridge between the PID and HTTP worlds that should be interoperable across PID systems:

- Given the PID of an asset, navigate to the HTTP-identified resources, other than the human-oriented landing page, that reside under the asset, for example the PDF or HTML version of a paper, the constituent files of a multi-file dataset, the provenance information for an asset, etc.
- Given the HTTP URI of a resource that resides under a scholarly asset, such as the resources mentioned in the previous bullet, determine the containing asset's PID.

In a web-based research communication environment dominated by human users, these needs were met through human interpretation. But applications and value added frameworks are rapidly emerging that require unambiguous machine-actionable approaches to meet those needs. Current PID frameworks have mostly focused on human use cases and have largely overlooked machine use.

PIDs and HTTP URIs: A Mapping

In order to support applications that use and add value to scholarly assets identified by PIDs, an unambiguous mapping is proposed between the PID-oriented paradigm and the HTTP-oriented web. The mapping is bi-directional as it provides:

- A uniform path from the `PID` of a compound scholarly asset to its constituent resources, each identified by a distinct `HTTP-URI-LOC`.
- A uniform path from the `HTTP-URI-LOC` of a constituent resource of a scholarly asset to the `PID` of that asset.

The essential ingredients of such a mapping, all rooted in existing standards and practice, are discussed in the remainder of this section. They are illustrated in Figures 1 and 2.

An interpretation of the nature of the resource identified by `HTTP-URI-PID`

It is interesting to observe that there is no common understanding of the nature of the resource identified by `HTTP-URI-PID`. Two possible interpretations immediately come to mind:

- *HTTP-URI-PID identifies the landing page associated with the asset identified by PID:* This interpretation is supported by the typical HTTP redirection (HTTP "302 Found") from `HTTP-URI-PID` to `HTTP-URI-LAND`.
- *HTTP-URI-PID identifies the asset identified by PID for the purpose of web interactions:* This interpretation is supported by the consideration that a scholarly asset is an intellectual object that belongs in the Work/Expression realm rather than in the Manifestation/Item realm of the FRBR hierarchy. In semantic web terms, this means that such a scholarly asset should be considered a non-information resource. This term is used to refer to resources that have no inherent digital representation, such as people, products, places, ideas and concepts. A *non-information resource* can have an HTTP URI to identify it on the web, in which case it is described by a document at a different HTTP URI typically reached through content negotiation. Recent DOI practice that makes descriptive RDF metadata pertaining to the asset identified by `PID` available through content negotiation with `HTTP-URI-PID` supports this interpretation. It is further supported by a CrossRef DOI Display Guideline that recommends using the `HTTP-URI-PID` to refer to the PID-identified asset in the online environment, overruling prior practice that consisted of using the `PID` for that purpose. This suggests equivalence between both identifiers, hence between what is being identified; with one identifier to be used for web interactions, the other for e.g. print.

Given the need for the availability of a machine-actionable document that describes a compound scholarly asset, the latter interpretation is used: the `HTTP-URI-PID` identifies the asset identified by `PID` for the purpose of web interactions.

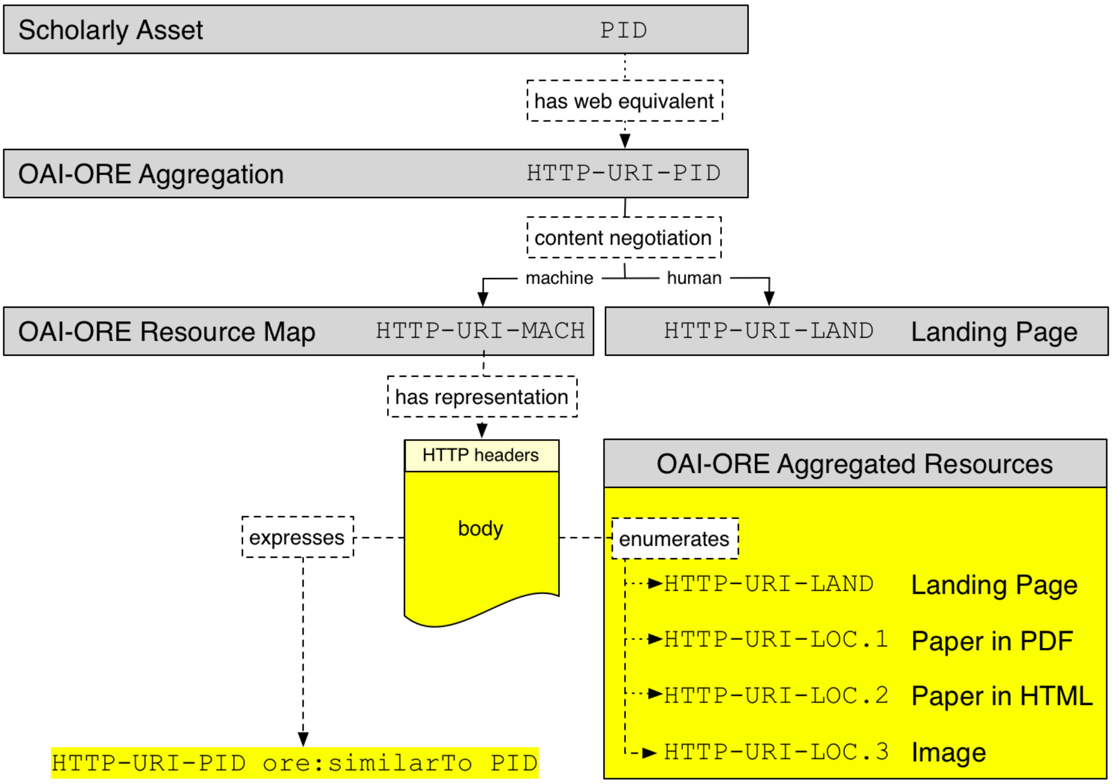

Figure 1: A path from the PID of a compound scholarly asset to HTTP URIs of its constituent resources

HTTP content negotiation with `HTTP-URI-PID`

Having classified the resource identified by `HTTP-URI-PID` as a *non-information resource*, best practice can be followed to obtain descriptions of that resource. Current practice is described in Cool URIs for the Semantic Web. Given the reality of the separation between the resolver of `HTTP-URI-PID` and the provider of associated content, the pattern 303 URIs Forwarding to Different Documents is appropriate. Under that approach, the resolver redirects (HTTP "303 See Also") to the landing-page (`HTTP-URI-LAND`) or to a machine-actionable description of the asset (`HTTP-URI-MACH`) depending on an agent's content negotiation preferences. It needs to be noted that this current practice remains a topic of heated discussions, and that eventually a different practice, which necessarily needs to provide the same functionality, may emerge.

A Landing Page for Machines: `HTTP-URI-MACH`

A landing page is introduced to adequately describe a compound scholarly asset in a machine-actionable manner. The description minimally enumerates the constituent resources of the asset and can further include metadata (descriptive, structural, technical, rights, provenance), types, and relationships pertaining to the asset itself as well as to each of its constituent resources. Given the description is targeted at machines, and interoperability across PID systems is desired, the use of RDF seems logical. The RDF-based, web-centric OAI-ORE specification, released in 2008, was specifically devised to describe of aggregations of web resources, and particularly for compound scholarly assets. The following are core characteristics when choosing the OAI-ORE approach:

- The resource identified by `HTTP-URI-PID` is regarded an OAI-ORE Aggregation.
- The resource identified by `HTTP-URI-MACH` is an OAI-ORE Resource Map that describes the Aggregation, and hence the compound scholarly asset, in a machine-actionable manner.
- The Resource Map `HTTP-URI-MACH` is available through content negotiation with `HTTP-URI-PID` in the manner described above.
- The constituent resources of the compound scholarly asset are Aggregated Resources of the Aggregation, each with a distinct `HTTP-URI-LOC`. They can be typed and related using terms from both cross-community and community-specific ontologies, such as done by SURF in The Netherlands under the umbrella "info-eu-repo". An Aggregated Resource can itself be an Aggregation, allowing support for compound scholarly assets that contain other compound scholarly assets, each with their own `PID`, `HTTP-URI-PID`, etc.
- Metadata pertaining to the compound asset can be expressed as an integral part of the Resource Map by means of RDF statements that have `HTTP-URI-PID` as their subject. It can also be provided as one or more appropriately typed Aggregated Resources. The latter approach was used in info-eu-repo with a bibliographic metadata resource typed as <info:eu-repo/semantics/descriptiveMetadata>.
- As a special metadata case, the `PID` is expressed by means of a statement that has `HTTP-URI-PID` as its subject and the `PID` as object. OAI-ORE introduces the special-purpose predicate <http://www.openarchives.org/ore/terms/similarTo> to that end. The use of this predicate requires the object of the RDF statement to be a resource,

and hence have a URI to identify it. Since none of the common PIDs (Handle, DOI, ARK) are in the URI Scheme Registry there is a need for a solution or convention with this regard. It needs to be noted that handles and DOIs can be expressed as URIs using the registered info URI scheme. For example, using the info URI scheme, the PID <doi:10.1145/1998076.1998111> is expressed as <info:doi/10.1145/1998076.1998111>

- The human landing page HTTP-URI-LAND can be treated as an Aggregated Resource, and typed accordingly. For example, in info-eu-repo it is typed as <info:eu-repo/semantics/humanStartPage>; the W3C Data Catalog vocabulary introduces the type <http://www.w3.org/ns/dcat#landingPage>.

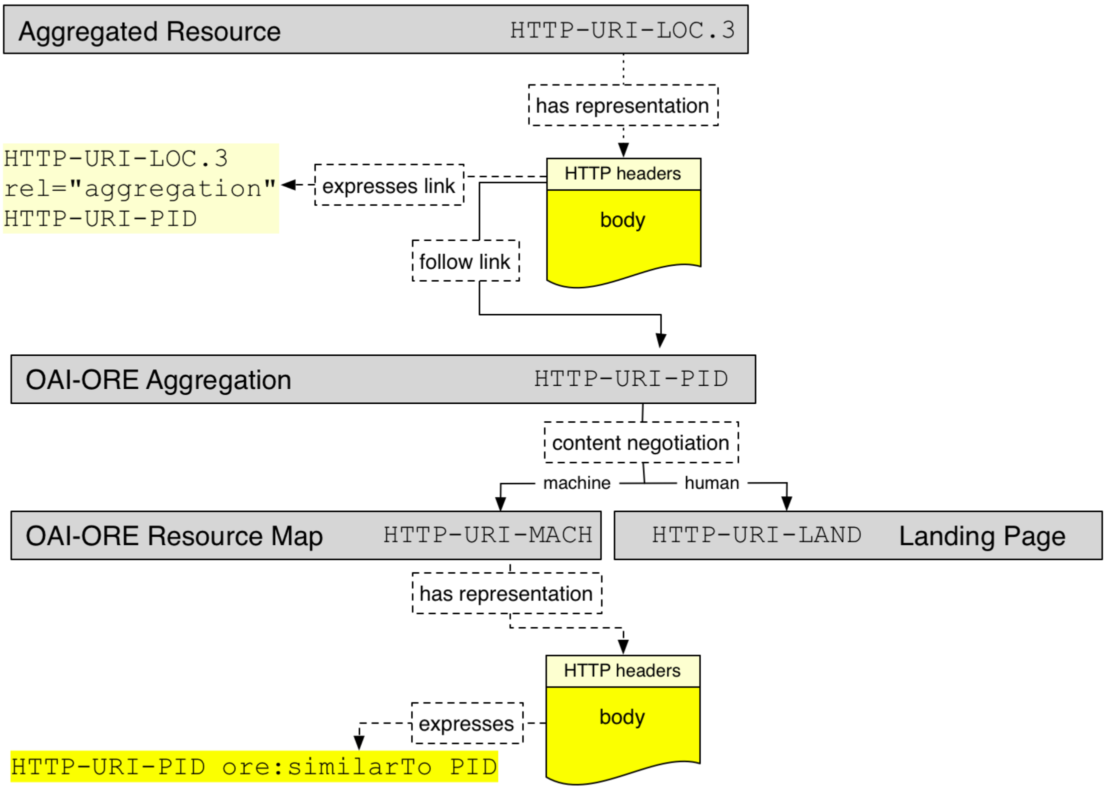

Figure 2: A path from a constituent resource of a compound scholarly asset, to the PID of that asset

HTTP Links

Figure 1 illustrates the path from the scholarly asset, identified by PID, to its constituent resources, each identified by a distinct HTTP-URI-LOC. To achieve the inverse path, HTTP Link headers, specified in RFC 5988 can be used.

When responding to HTTP requests issued against the HTTP-URI-LOC of an Aggregated Resource, an HTTP Link header is included containing a link that relates this resource to the Aggregation it resides under. To that end, the OAI-ORE specification introduces the “aggregation” relation type specifically intended to point to OAI-ORE Aggregations. However, recent practice suggests a preference for using more generic relation types combined with the use of a media type that indicates further application details. Following that practice, the generic “collection” relation type, registered in the IANA Relation Type Registry, could be used to point to an Aggregation, and a media type for Resource Maps would need to be agreed upon. The latter would also be beneficial for the purpose of content negotiation with HTTP-URI-PID, as it would allow requesting different types of RDF metadata, e.g. descriptive, Resource Map. It needs to be noted that a resource can reside under multiple Aggregations. In that case, multiple links should be provided in the HTTP Link header, one pointing to each Aggregation. Also, an Aggregated Resource can be hosted by a different system than the Aggregation it resides under. This can, for example, be the case when a special-purpose archive hosts a dataset and a publisher hosts the paper that builds on it. In that case, a pingback mechanism, such as WebMention, can be used to inform a resource that it was aggregated.

After having received an HTTP response from an HTTP-URI-LOC, a client can follow the link to an Aggregation. As described above, content negotiation with the Aggregation’s HTTP-URI-PID then leads to a Resource Map, which includes an RDF statement that relates the HTTP-URI-PID to the PID by means of the <http://www.openarchives.org/ore/terms/similarTo> predicate, if there is a URI form of the PID available. As such, the HTTP-URI-LOC of the Aggregated Resource is mapped back to the PID of the scholarly asset it belongs to. Note that RDF statement that relates the HTTP-URI-PID to the PID can also be expressed as a link in the HTTP Link header in responses to HTTP requests for the Aggregation’s HTTP-URI-PID. This link has HTTP-URI-PID as Context IRI, < http://www.openarchives.org/ore/terms/similarTo> as relation type and PID, expressed according to a URI scheme, as Target IRI. Inclusion of this link removes the need to request the Resource Map to obtain the PID of the scholarly asset.

Resource Versioning

On the Web, a common resource versioning pattern exists that consists of:

- Having a *generic URI* where at any moment in time the current version of the resource is accessible.
- Having a dedicated *version URI* for each resource version.

This versioning pattern can, for example, be observed on top of W3C Specifications such as the Architecture of the World Wide Web:

- <http://www.w3.org/TR/webarch/> is the generic URI for that specification;
- <http://www.w3.org/TR/2004/REC-webarch-20041215/> is the version URI for the current version of that specification;
- <http://www.w3.org/TR/2004/PR-webarch-20041105/> is the version URI for the previous version of that specification;
- <http://www.w3.org/TR/2004/WD-webarch-20040816/> is the version URI for the version before that; etc.

As described in Resource Versioning and Memento, this common versioning pattern aligns nicely with the Memento “Time Travel for the Web” protocol specified in RFC

7089. The protocol can be supported in a modular manner to achieve one or more of the following functionalities related to discovering and accessing resource versions:

- Expressing the datetime of a resource version: This is achieved by using the Memento-Datetime HTTP header when responding to HTTP HEAD/GET requests issued against a version URI.
- Interlinking resource versions: This is achieved by using an HTTP Link header when responding to HTTP HEAD/GET requests issued against a version URI. The Link header includes links with the "memento" relation type that point at various version URIs; each link includes the datetime of the target resource version as a link attribute.
- Providing an overview of resource versions: This is achieved by using an HTTP Link header when responding to HTTP HEAD/GET requests issued against a version URI. The Link header includes a link with the "timemap" relation type that points at a TimeMap, a document that lists all version URIs and and their associated version datetime.
- Negotiating in time with the generic URI to obtain the version URI that was the current one at a specified moment in time: This is the most powerful feature of the Memento protocol. It is triggered by including an Accept-Datetime header with the desired version datetime as value when issuing an HTTP HEAD/GET against the generic URI.

Since they are identified by means of HTTP URIs, Aggregations, Resource Maps, and Aggregated Resources including landing pages can use the described versioning pattern. This requires the use of versioning systems to manage these resources. A trend in that direction can be observed in research communication, as exemplified by the increased use of wikis and datawikis, the growing interest in versioning file systems and databases that can be rolled back to prior states.

In the PID paradigm, a distinct PID is typically assigned per version of an asset. As a result, the bridge can work as described per asset version by having a distinct version URI per Aggregation version, Resource Map version, and Aggregated Resource versions. The first two Memento related features listed above can also be implemented because they only entail version URIs and version datetimes.

It is less common in the PID paradigm to assign a cross-version PID, which, in a manner comparable to the generic URI approach, at any point in time refers to the current version of the asset. If such a *generic PID* would be introduced, with its associated *generic HTTP-URI-PID*, *generic HTTP-URI-LAND*, and *generic HTTP-URI-MACH*, the remaining two features of the Memento protocol - the list of versions and datetime negotiation - can also be supported. The latter allows navigating towards a description of the compound scholarly asset for human and machine consumption, as it existed at a specified time in its lifecycle.

Conclusion

The proposed bridge would make the research communication environment more machine-friendly and can pave the way for new value-added applications. As an example, in the case of scholarly annotation as envisioned by the web-oriented, RDF-based, W3C Open Annotation Community Group effort, the proposed bridge provides answers to the following rather fundamental questions:

- *Which HTTP URI should be used for the Target of an annotation when that annotation pertains to a PID-identified asset?* By interpreting `HTTP-URI-PID` as equivalent to `PID` for the purpose of web transactions, it follows that `HTTP-URI-PID` should be used as the Target URI for such annotations.
- *How can annotations that pertain to a PID-identified asset trickle down, if so desired, to the resources that reside under that asset?* Collecting annotations that pertain to a resource `HTTP-URI-LOC` translates to discovering, for example by means of a SPARQL query, which annotations have `HTTP-URI-LOC` as Target of the annotation. If it is desired to additionally inherit annotations made to the asset under which `HTTP-URI-LOC` resides, the discovery operation needs to include annotations that have `HTTP-URI-PID` as Target. As described, doing an HTTP HEAD on `HTTP-URI-LOC` and parsing the link to the Aggregation from the HTTP Link header yields the desired URI.
- *How can annotations that are made to a resource that resides under a PID-identified asset trickle back up to the asset, if so desired?* Collecting annotations that pertain to the asset translates to discovering which annotations have `HTTP-URI-PID` as Target of the annotation. If it is desired to additionally inherit annotations made to resources that reside under the asset, the discovery operation needs to include annotations that have the `HTTP-URI-LOC`s of Aggregated Resources as Target. As described, these URIs are enumerated in the Resource Map `HTTP-URI-MACH` that is available through content negotiation with `HTTP-URI-PID`.

The bridge is based on existing standards and practice: HTTP, Cool URIs for the Semantic Web, HTTP Links, IANA relation types, OAI-ORE, Memento. It can achieve interoperability across PID systems by establishing interoperability between each PID system and the web and its technology stack. Two issues were highlighted that need to be addressed:

- Expressing PIDs as URIs so they can be used as objects in RDF statements and as Target IRIs in HTTP Links.

- In case the “collection” rather than “aggregation” relation type would be used in HTTP Links to point to an Aggregation, which media type should be assigned to Resource Maps?

In addition, the more involving task presents itself of deciding which ontologies to use for expressing metadata, types, and relationships in RDF Resource Maps in order to achieve a satisfactory level of semantic interoperability for research communication. With that regard, the Linked Open Data mantra Reuse Vocabularies Whenever Possible should be followed and the list of common vocabularies provided by State of the LOD Cloud can serve as inspiration. The reuse of common vocabularies can lead to a coarse-grained cross-community profile of Resource Maps that would also help to integrate research communication into the Linked Data cloud. The described tasks are modest but do require input from all constituents of the research communication endeavour.

Acknowledgments

Thanks for feedback and inspiration to: Sean Bechhofer, Maarten Hoogerwerf, Pete Johnston, Carl Lagoze, Michael L. Nelson, Andrew Treloar, Simeon Warner